GESIS Leibniz Institute for the Social Sciences

Enhancing **Software-Related Information Extraction** via **Single-Choice Question Answering** with **Large Language Models**

Wolfgang Otto, Sharmila Upadhyaya, Stefan Dietze

Shared Task: Software Mention Detection in Scholarly Publications (**SOMD**)

1. Identify Mentions of 5 Software related entity types incl. Intention (e.g., **Application, Plugin, Operating System**)
2. Detect 10 attributive entities (e.g., **Version, License, Developer**)
3. Identify 13 typed relations between the Mentions (e.g., **version of**, **license of**)

Sentence:

"This example file runs on BEAST (Application) 2.1.0 (Version) *with the* MultiTypeTree (Plugin) *package (http://compevol.github.io/MultiTypeTree/ URL)."*

Motivation

- Tracking Software Usage
- Foundation for Reproducablility Studies
- Foster Transparency in Research

Given Dataset

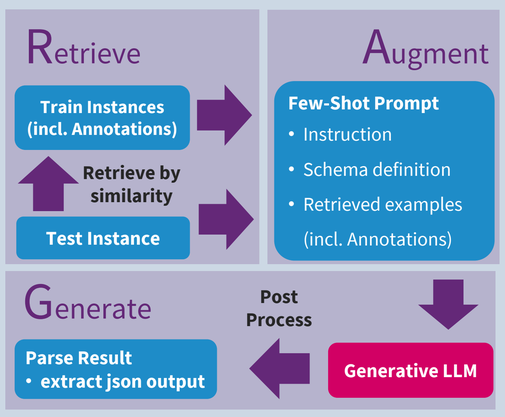
- ~1,400 Publication
- ~48,000 Sentences
- ~3,800 Software Mentions
- Bio and medical domain (Pubmed + PLoS)
- Based on SoMeSci
- used for SoftwareKG

	Train	Test	Test share
All sentences	39,769 (100.0%)	8181 (100.0%) ST1 Data	17.1%
Incl. entity annotations (ST1 Task)	2,326 (5.8%)	373 (4.6%) ST2 Data	13.8%
Incl. attributive annotations (ST2 Task)	1,690 (4.3%)	230 (2.8%)	12.0%
Incl. relation annotations (ST3 Task)	1,075 (2.7%)	128 (1.6%) ST3 Data	10.6%

Approach

- Show expected scope of task
- Show similar entities/relations
- Show similar contexts
- **Idea**: Use Training Contexts as examples

Retrieval-Augmented Generation for NER



Sentence:

*"This example file runs on **BEAST 2.1.0** with the **MultiTypeTree** package [..]. "*

Example Question:

Which of the statements about the Version '2.1.0' is true?

☑ *'2.1.0' is the version of 'BEAST'.*

☐ *'2.1.0' is the version of 'MultiTypeTree'.*

Derived Relation: 2.1.0 (Version) — version of → BEAST (Application)

Relation Extraction as Question Answering

Observations:

- Every attributive entity has exactly one related entity
- If there is only one candidate entity in a sentence we know the answer (i.e., a "necessary" relation)
- For the others, we need to decide between a few (cf. Single Question Answering)

→ *For relations with multiple candidate, we reformulate the RE Task as a single-choice QA task*

Results

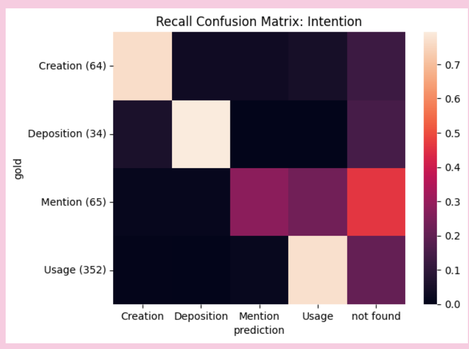
Retrieval Examples: Entity Similarity

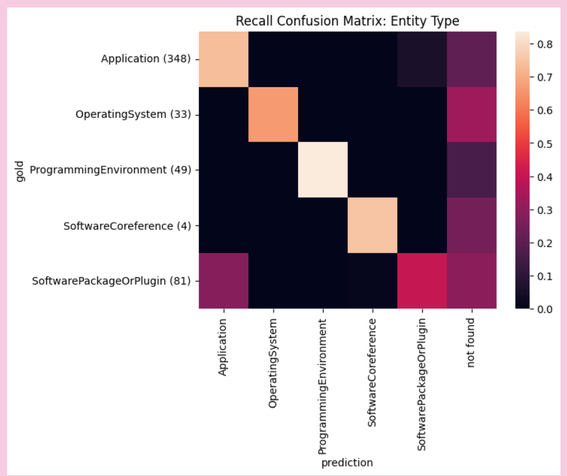
entity	label	sim	split	sentence
PhosphOrtholog	Application Creation	1.00	test	To this end , we have developed an automated web - based tool , PhosphOrtholog , which allows batch processing and mapping of large species - specific PTM datasets to compare overlap at a site - specific level .
SNPdetector	Application Creation	0.93	train	We developed a software tool , SNPdetector , for automated identification of SNPs and mutations in fluorescence - based resequencing reads .
ESPRIT - Forest	Application Creation	0.92	train	In this paper we developed a new algorithm called ESPRIT - Forest for parallel hierarchical clustering of sequences .

Retrieval Examples: Sentence Similarity

split	sim	text
test	1.00	To this end , we have developed an automated web - based tool , PhosphOrtholog , which allows batch processing and mapping of large species - specific PTM datasets to compare overlap at a site - specific level .
train	0.99	Here we present a tool , Podbat (Positioning database and analysis tool) , that incorporates data from various sources and allows detailed dissection of the entire range of chromatin modifications simultaneously .
train	0.99	We designed and developed a new method , MSACompro , to synergistically incorporate predicted secondary structure , relative solvent accessibility , and residue - residue contact information into the currently most accurate posterior probability - based MSA methods to improve the accuracy of mult...
train	0.99	In this paper , we present a lossless compression tool , MAFCO , specifically designed to compress MAF (Multiple Alignment Format) files .

NER Recall Confusion





RAG Performance for Task 1 (NER)

Paradigm	Retrieval	F1	Model	parameter
Finetuned	-	0.599	SciBERT	-
Prompt	Random	0.483	GPT 3-5	Random k=7
Prompt	Random	0.525	GPT 3-5	Random all entity types shown
Prompt	Sim. sentences	0.647	GPT 3.5	topn=10
Prompt	Sim. entities	0.624	GPT 3.5	topn=7 x n entities
Prompt	Random	0.574	GPT 4	Random all entity types shown
Prompt	Sim. sentences	0.677	GPT 4	topn=10
Prompt	Sim. entities	**0.679**	GPT 4	topn=7 x n entities

- #1 System for **SOMD** Subtask 2 and 3
- Entity similarity catch similar entities **and** similar context
- Best on NER for low data regime
- Single-Choice QA for RE: best performing Approach

SOMD Results + RE Baselines

Task	#	User	F1	Precision	Recall	Task	#	User	F1	Precision	Recall
1	1	**phinx**	**.740**	**.761**	**.750**	2	1	***ours***	***.838***	***.835***	***.847***
1	2	david-s477	.692	.739	.711	2	2	phinx	.743	.745	.748
1	3	ThuyNT03	.678	.729	.649	3	1	***ours***	***.916***	*.911*	***.924***
1	4	*ours*	*.652*	*.679*	*.664*	3	2	phinx	.897	.900	.897
1	5	vampire	.648	.682	.637	3	-	*baseline*	*.864*	*.857*	*.875*
						3	-	*necessary*	*.562*	***.933***	*.415*

Contact: Wolf Otto (wolfgang.otto@gesis.org)

Heinrich Heine Universität Düsseldorf
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
In cooperation with nfdi
BERD @NFDI
Leibniz Association